SAIN TEKOM
ISSN 2088-1770 (Print)
ISSN 2503-3247 (Online)
Vol. 15 No.1 (2025) 15-28

JURNAL
SAINTEKOM
Sains, Teknologi, Komputer, dan Manajemen

Perbandingan Algoritma Machine Learning Menggunakan Pemilihan Fitur Chi-square dalam Pengklasifikasian Penyakit Jantung

*Hirmayanti[1], Ema Utami[2]
[1,2)]Magister Teknik Informatika, Universitas Amikom Yogyakarta
Jl. Ring Road Utara, Condong Catur, Sleman, Yogyakarta
Email: [1]hirmayanti@students.amikom.ac.id, [2]ema.u@amikom.ac.id

ABSTRACT

Heart disease is one of the deadliest diseases worldwide. This condition often presents symptoms that do not immediately cause severe effects on the sufferer, making early anticipation crucial. To reduce fatalities caused by heart disease or cardiovascular disorders, a system is required to identify its primary causes so that these factors can be minimized. Therefore, this study applies the Chi-square feature selection method to determine the key features influencing the accuracy of Machine Learning models. A comparison is conducted between K-Nearest Neighbor, Naïve Bayes, Logistic Regression, Support Vector Machine, and Random Forest algorithms. This comparison aims to obtain the most accurate results, as a higher algorithm accuracy leads to a more precise classification system for heart disease. The study's findings indicate that eight key features selected using the Chi-square method yield the highest accuracy, specifically 93.51% with the KNN algorithm. These results demonstrate that using relevant features improves classification accuracy and system efficiency compared to utilizing all available features. Consequently, this research contributes to the selection of essential features in Machine Learning algorithms through the Chi-square technique, ensuring a more effective and optimized heart disease classification system.

Keywords : heart disease; cardiovascular; feature selection; chi-square; hyperparameter

ABSTRAK

Penyakit jantung termasuk penyakit yang mematikan di seluruh dunia. Penyakit ini seringkali gejalanya tidak langsung memberikan dampak yang begitu parah terhadap si penderita, oleh karena itu sangat perlu untuk diantisipasi. Untuk mengurangi korban akibat penyakit jantung atau *cardiovasculas* ini, dibutuhkan adanya sistem yang mampu mengidentifikasi penyebab utama dari penyakit ini sehingga faktor atau penyebab tersebut bisa diminimalisir. Oleh karena itu, pada penelitian ini menggunakan *feature selection Chi-square* untuk memilih fitur utama yang berpengaruh terhadap akurasi model *Machine Learning*, dengan membandingkan algoritma *K-Nearest Neighbor, Naïve Bayes, Logistic Regression, Support Vector Machine*, dan *Random Forest*. Perbandingan ini dilakukan untuk memperoleh hasil yang akurat, semakin tinggi akurasi algoritma maka semakin tinggi pula keakuratan sistem yang dihasilkan dalam mengklasifikasikan penyakit jantung. Berdasarkan hasil penelitian, diketahui bahwa ada 8 fitur utama dari *Chi-square* yang menghasilkan akurasi tertinggi yaitu 93.51% dari algoritma KNN. Berdasarkan hasil penelitian ini, penggunaan fitur relevan atau utama mampu menghasilkan sistem yang akurat dan efisien dalam mengklasifikasikan penyakit jantung, bila dibandingkan dengan menggunakan semua fitur. Oleh karena itu, penelitian ini berkontribusi pada pemilihan fitur penting dalam algoritma *Machine Learning* melalui teknik *Chi-square*, yang memastikan sistem klasifikasi penyakit jantung yang lebih efektif dan optimal.

Kata kunci : penyakit jantung; *cardiovascular;* pemilihan fitur; *chi-square; hyperparameter*

Diterima Redaksi: 22-1-2025 | Selesai Revisi: 11-2-2025 | Diterbitkan Online: 31-3-2025

1. PENDAHULUAN

Penyakit jantung merupakan salah satu penyakit yang banyak memakan korban di seluruh dunia (Khan et al., 2023). Diperkirakan sekitar 17,9 juta orang yang menderita penyakit *cardiovascular* berdasarkan data dari WHO (World Health Organization, 2021). Penyakit jantung atau *cardiovascular* tidak langsung menyerang atau memberikan rasa sakit terhadap si penderita, namun tanpa disadari penyakit ini banyak memakan korban baik itu laki-laki maupun perempuan. Banyak faktor pemicu penyakit ini seperti merokok, obesitas, diabetes, tekanan darah tinggi, kurangnya aktivitas bergerak, kadar lemak darah tidak normal, genetik bahkan juga pemicunya dari gaya hidup sehari-hari (Yahaya et al., 2020).

Beberapa penelitian seperti (Khan et al., 2023), (Ozcan & Peker, 2023) dan (Bhatt et al., 2023) telah melakukan penelitian mengenai prediksi penyakit jantung. Berbagai metode digunakan untuk mendapatkan hasil evaluasi yang akurat, guna membangun sistem yang bisa mendiagnosis penyakit jantung secara tepat dan efisien.

Penelitian (Chandrasekhar & Peddakrishna, 2023) menerapkan metode *ensemble*, sehingga diperoleh hasil lebih akurat bila dibandingkan dengan evaluasi model klasik. Namun penelitian ini belum bisa menentukan fitur utama yang berpengaruh terhadap nilai akurasi model, sehingga masih dibutuhkan pengembangan lagi.

Penelitian (Albert et al., 2023) juga melakukan penelitian untuk mendiagnosis penyakit jantung dengan mengusulkan model BOML (*Balanced and Optimized Machine Learning*) menggunakan algoritma CART. Penelitian ini menggunakan CART untuk mengidentifikasi fitur yang berpengaruh terhadap prediksi penyakit jantung. Namun penelitian ini lebih berfokus pada peningkatan akurasi dengan membandingkan berbagai metode *balanced* seperti *unsampled, smote, adasyn, borderline smote, rose* dan *safe level smote.*

Dengan berfokus pada penggunaan fitur yang tepat seringkali menghasilkan sistem yang lebih akurat dan efisien, bila dibandingkan dengan penggunaan semua fitur (Rao & Srivastava, 2024). Oleh karena itu, dibutuhkannya pemilihan fitur utama

DOI: https://doi.org/10.33020/saintekom.v15i1.815

yang akan digunakan pada proses pengklasifikasian model. Pemilihan fitur secara manual dapat menghilangkan fitur yang relevan, jika tidak dilakukan oleh ahlinya serta dibutuhkan ketelitian dalam menentukan fitur yang akan digunakan guna meminimalisir kesalahan (Adiatma et al., 2021). Oleh sebab itu, pada penelitian ini akan menggunakan teknik *feature selection* untuk mendapatkan fitur yang relevan berdasarkan perhitungan statistik *Chi-square* (Chi2).

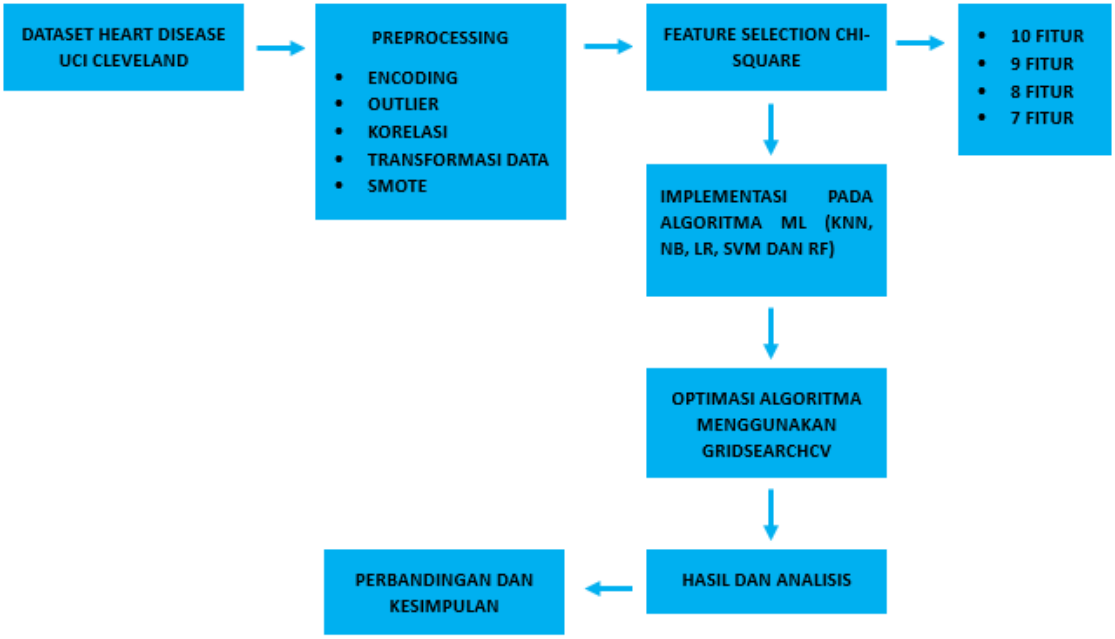
Pemilihan fitur menggunakan *Chi-square* pada penelitian sebelumnya telah banyak dilakukan perbandingan dengan teknik *feature selection* lainnya dan diperoleh *Chi-square* unggul (Sarra et al., 2022), (Ogundepo & Yahya, 2023), (Biswas et al., 2023), (Khurana et al., 2021), (Spencer et al., 2020) dan (Reddy et al., 2021).

Untuk membatasi penelitian ini supaya tetap pada pembahasannya dan tidak melebar kemana-mana, maka peneliti memberikan batasan yaitu penelitian ini menggunakkan teknik pemilihan fitur *Chi-square* (Chi2) untuk memilih fitur yang relevan, algoritma yang digunakan merupakan algoritma klasifikasi meliputi KNN (Jusia et al., 2024), NB (Spencer et al., 2020), LR (Estetikha et al., 2021), SVM (Li et al., 2020), RF (Sushma et al., 2021)(Biswas et al., 2023). Untuk menyempurnakan penelitian sebelumnya, penelitian ini menggunakan teknik optimasi algoritma yaitu *GridSearchCV* (Claesen & De Moor, 2015). Hal tersebut dilakukan guna mengoptimalkan kinerja dari setiap algoritma yang digunakan, dan programnya dijalankan menggunakan *Google Collab* dengan bahasa Python.

2. METODE

Pada penelitian ini menggunakan dataset *heart disease* Cleveland yang diambil dari Kaggle. Sebelum diterapkan pada algoritma, ada beberapa tahapan yang dilakukan guna menyiapkan dataset yang disebut *preprocessing*. Adapun *preprocessing* yang dilakukan meliputi *encoding*, *outlier*, korelasi, transformasi data dan SMOTE. Untuk lebih jelasnya, tahapan pada penelitian ini bisa dilihat pada Gambar 1.

DOI: https://doi.org/10.33020/saintekom.v15i1.815



Gambar 1. Alur Penelitian

2.1. Dataset

Pada penelitian ini menggunakan dataset *heart disease* yang diambil dari UCI *Machine Learning Repository*, di mana dataset tersebut memiliki total data sebanyak 303 data dan 14 fitur. Untuk dataset tersebut bisa diakses pada https://archive.ics.uci.edu/dataset/45/heart+disease.

2.2. Preprocessing

Langkah selanjutnya pada penelitian ini adalah menyiapkan dataset sebelum digunakan atau disebut dengan *preprocessing*. Adapun preprocessing yang dilakukan pada penelitian ini meliputi *encoding*, mengatasi *outlier*, pengecekan korelasi, transformasi dan menyeimbangkan dataset menggunakan SMOTE.

Dataset ini tidak ditemukan adanya *missing value*, namun ditemukan adanya atribut yang memiliki *value* objek seperti pada Gambar 2.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       303 non-null    float64
 1   sex       303 non-null    float64
 2   cp        303 non-null    float64
 3   trestbps  303 non-null    float64
 4   chol      303 non-null    float64
 5   fbs       303 non-null    float64
 6   restecg   303 non-null    float64
 7   thalach   303 non-null    float64
 8   exang     303 non-null    float64
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    float64
 11  ca        303 non-null    object
 12  thal      303 non-null    object
 13  num       303 non-null    int64
dtypes: float64(11), int64(1), object(2)
memory usage: 33.3+ KB
```

Gambar 2. Tipe Data

DOI: https://doi.org/10.33020/saintekom.v15i1.815

Berdasarkan Gambar 2 diketahui hasil pengecekan dari tipe data dan ditemukan adanya atribut yang bernilai objek, dan atribut yang bernilai objek (selain dari numerik) tidak bisa dibaca oleh komputer. Oleh karena itu, peneliti menggunakan metode *encoding*. *Encoding* merupakan teknik yang digunakan untuk mengubah nilai kategori menjadi nilai numerik (Poslavskaya & Korolev, 2023). Sedangkan *outlier* adalah nilai atau data dari atribut yang menyimpang dari data kebanyakan lainnya (Escalante, 2005).

Kelas target (*num*) pada penelitian ini memiliki *multi-class* seperti yang bisa dilihat pada Gambar 3, sehingga peneliti melakukan transformasi data dari yang bernilai *multi-class* menjadi *binary class* seperti pada Gambar 4.

num	count
0	164
1	55
2	36
3	35
4	13

dtype: int64

Gambar 3. Dataset dengan *Value Multi-class*

Pada Gambar 3 menunjukkan bahwa pada dataset yang digunakan terdapat lima label diagnosa penyakit jantung, di mana 0 menunjukkan diagnosis normal dan 1-4 terdeteksi terkena penyakit jantung (1 : ringan, 2 : sedang, 3 : parah, 4 : sangat parah). *Multi-class* merupakan cara untuk menginterpretasikan data yang diklasifikasikan ke dalam tiga atau lebih kategori. Pada penelitian ini, peneliti melakukan transformasi *multi-class* menjadi *binary class* untuk mengurangi redudansi data. *Binary class* merupakan interpretasi data yang diklasifikasikan menjadi dua kategori yang terdiri dari 0 dan 1 (Sokolova & Lapalme, 2009).

num	count
0	219
1	84

dtype: int64

Gambar 4. Dataset dengan *Value Binary Class*

Dengan menjadikan kelas target (*num*) menjadi *binary class*, mesin akan lebih mudah menerima masukkan sehingga proses pengklasifikasian lebih efektif (Kibria & Matin, 2022). Diketahui dataset ini memiliki nilai yang tidak seimbang pada kelas target, di mana untuk kelas 0 berjumlah 219 baris dan kelas 1 berjumlah 84 baris. Oleh karena itu, peneliti menggunakan metode SMOTE untuk menyeimbangkan data. Dan untuk pembagian datasetnya digunakan *set training* dan *set testing* dengan rasio 7:3.

DOI: https://doi.org/10.33020/saintekom.v15i1.815

SMOTE merupakan teknik untuk menangani ketidakseimbangan data berdasarkan algoritma *random sampling* (Bujang et al., 2021). Peneliti menggunakan SMOTE karena pada penelitian sebelumnya SMOTE telah menunjukkan kinerja bagus dalam menangani ketidakseimbangan data (Narayanan & Jayashree, 2024)(Yulianto et al., 2024).

2.3. Chi-square

Pada pemilihan fitur menggunakan *Chi-square*, Chi2 akan melakukan penyeleksian fitur berdasarkan statistik *Chi-square* yang bekerja dengan cara mencari selisih diantara frekuensi yang diamati dan frekuensi yang diharapkan, kemudian dikuadratkan dan membaginya dengan frekuensi yang diharapkan (Plackett, 1984). Untuk lebih jelasnya bisa dilihat pada persamaan (1).

$$\boldsymbol{Chi2} = \frac{(\boldsymbol{O_i - E_i})}{\boldsymbol{E_i}} \quad (1)$$

Diketahui : Oi adalah frekuensi yang diamati dan Ei adalah frekuensi yang diharapkan.

Tabel 1. Rangking Fitur Berdasarkan *Chi-square*

Rangking	Fitur	Rangking	Fitur
1	*oldpeak*	8	*sex*
2	*thalach*	9	*trestbps*
3	*ca*	10	*chol*
4	*exang*	11	*restecg*
5	*cp*	12	*fbs*
6	*slope*	13	*age*
7	*thal*		

2.4. Algoritma Klasifikasi

Penelitian ini menggunakan dan membandingkan berbagai algoritma klasifikasi *Machine Learning* seperti K-*Nearest Neighbor* (KNN), *Naïve Bayes* (NB), *Logistic Regression* (LR), *Support Vector Machine* (SVM) dan *Random Forest* (RF). Setelah dataset dibersihkan pada tahap *preprocessing*, selanjutnya dataset diterapkan pada model. Penggunaan algoritma-algoritma ini bertujuan untuk menemukan model yang tepat dan akurat untuk melakukan klasifikasi penyakit jantung. Algoritma *Logistic Regression* pada penelitian (G et al., 2022) memprediksi penyakit *cardiovascular* dengan melakukan pengujian terhadap jumlah data dan mencapai akurasi tertinggi sebesar 87.10%.

Algoritma KNN menunjukkan kinerja yang bagus pada penelitian (Khairi et al., 2021), penelitian Khairi dkk. menguji kinerja algoritma KNN sebanyak tiga kali yaitu dengan k = 3, k = 5 dan k = 7. Adapun hasil akurasi

DOI: https://doi.org/10.33020/saintekom.v15i1.815

tertinggi ditunjukkan dari $k = 5$ dan $k = 7$ yaitu 98.68%. Untuk penggunaan KNN dari segi nilai *k*, apabila KNN digunakan untuk klasifikasi maka nilai *k* harus ganjil, namun jika digunakan untuk prediksi maka nilai *k* dapat berupa bilangan ganjil maupun genap.

Begitupun dengan algoritma Naive Bayes yang memperoleh nilai akurasi lebih unggul dibandingkan algoritma lainnya dalam memprediksi penyakit jantung yaitu dengan akurasi 85% (Spencer et al., 2020). Algoritma *Naive Bayes* merupakan algoritma yang cukup populer dengan teorema bayes dan dikenal dengan algoritma yang mudah dan cepat dalam melakukan prediksi maupun pengklasifikasian terhadap data uji.

Algoritma SVM bekerja dengan mencari *hyperplane* terbaik yang memisahkan dua kelas data, setelah *hyperplane* terbaik ditemukan maka SVM dapat digunakan untuk mengklasifikasikan data yang baru. Pada penelitian (Li et al., 2020) melakukan pengklasifikasian penyakit jantung memperoleh SVM bekerja dengan baik dengan akurasi 92.37%. Sedangkan algoritma *Random Forest* telah banyak digunakan karena algoritma ini terdiri dari beberapa pohon keputusan yang dibuat secara acak guna menghasilkan prediksi yang akurat. Dan itu dibuktikan pada penelitian (Sushma et al., 2021) yang melakukan penelitian untuk mendeteksi penyakit jantung dan didapatkan akurasi tertinggi yaitu 99.3%.

2.5. Matrix Confusion

Untuk mengetahui seberapa akurat hasil pengujian pada penelitian ini, peneliti menggunakan *matrix confusion* sebagai evaluasi model. *Matrix confusion* merupakan evaluasi model yang banyak digunakan terhadap berbagai macam penelitian seperti (Shaon et al., 2024), (Spencer et al., 2020), (Li et al., 2020), (Sushma et al., 2021), (Ghosh et al., 2021). *Matrix confusion* dalam mengevaluasi sistem memiliki komponen seperti *Accuracy*, *Precision*, *Recall* dan *F1 score*.

Accuracy adalah jumlah nilai benar secara keseluruhan yang diperoleh dari data pengujian algoritma. *Precision* adalah nilai hasil evaluasi dari model yang membuat prediksi yang benar untuk kelas positif dari total prediksi positif. *Recall* adalah nilai hasil evaluasi dari model dalam mengidentifikasi kelas positif dengan benar. Sedangkan F1

DOI: https://doi.org/10.33020/saintekom.v15i1.815

score adalah kombinasi dari *precision* dan *recall* (Sokolova & Lapalme, 2009).

Untuk rumus perhitungan *Accuracy* bisa dilihat pada persamaan (2), *precision* pada persamaan (3), *recall* pada persamaan (4) dan *F1 score* pada persamaan (5).

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TP+FP+FN}} \quad (2)$$

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP\_}} \quad (3)$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \quad (4)$$

$$\boldsymbol{F1\ score} = \boldsymbol{2x}\frac{\boldsymbol{RecallxPrecision}}{\boldsymbol{Recall+Precision}} \quad (5)$$

Diketahui :

TP adalah *True Positives,* jumlah prediksi yang benar dalam memprediksi yang terdiagnosis penyakit jantung dan sesuai dengan kenyataannya.

TN adalah *True Negatives,* jumlah prediksi yang benar dalam memprediksi yang tidak terdiagnosis penyakit jantung (normal) dan sesuai dengan kenyataannya.

FP adalah *False Positives,* jumlah prediksi yang benar dalam memprediksi yang terdiagnosis penyakit jantung, namun kenyataannya salah.

FN adalah *False Negatives,* jumlah prediksi yang benar dalam memprediksi yang tidak terdiagnosis penyakit jantung (normal), namun kenyataannya salah.

3. HASIL DAN PEMBAHASAN

3.1. Hasil

Pada penelitian ini menggunakan teknik *Chi-square* (Chi2) untuk memilih fitur relevan, di mana pada penelitian ini melakukan pengujian sebanyak lima kali yaitu pengujian dengan semua fitur, 10 fitur, 9 fitur, 8 fitur dan 7 fitur. Hal ini dilakukan untuk melihat seberapa kuat pengaruh jumlah fitur yang digunakan dalam melakukan pengklasifikasian penyakit jantung. Sebelum diimplementasikan pada model, dataset dibagi menjadi 70% set latihan dan 30% set pengujian. Pada Gambar 5 menampilkan hasil evaluasi menggunakan kelima model, dan diperoleh akurasi tertinggi diperoleh oleh algoritma *Naive Bayes*, *Logistic Regression* dan *Random Forest* yaitu 75.82%.

	Model	Accuracy	Precision	Recall	F1
3	GaussianNB	75.824176	87.344086	3.956069	0.773077
0	Logistic Regreesion	75.824176	86.677419	3.745410	0.761538
4	Random Forest	75.824176	92.225806	4.591297	0.715385
1	SVM	74.725275	91.892473	3.306937	0.730769
2	KNeighbors	69.230769	88.666667	5.022599	0.738462

Gambar 5. Hasil Evaluasi Menggunakan Semua Model

Sedangkan untuk hasil evaluasi menggunakan *Chi-square* dengan 10 fitur relevan bisa dilihat pada Gambar 6.

DOI: https://doi.org/10.33020/saintekom.v15i1.815

	Model	Accuracy	Precision	Recall	F1
0	Logistic Regreesion	79.120879	87.354839	4.844727	0.796154
2	KNeighbors	78.021978	91.569892	5.622782	0.800000
3	GaussianNB	78.021978	88.000000	4.308113	0.800000
1	SVM	76.923077	91.580645	5.607514	0.746154
4	Random Forest	74.725275	91.913978	5.797644	0.696154

Gambar 6. Hasil Evaluasi Menggunakan 10 Fitur Relevan

Berdasarkan Gambar 6, hasil evaluasi menggunakan 10 fitur relevan dari *Chi-square* diperoleh nilai akurasi tertinggi dihasilkan dari *Logistic Regression* sebesar 79.12%. Bila dibandingkan dengan akurasi LR menggunakan semua fitur, diperoleh peningkatan 3.3% dengan 10 fitur relevan.

	Model	Accuracy	Precision	Recall	F1
2	KNeighbors	78.021978	89.935484	6.024231	0.800000
3	GaussianNB	78.021978	88.311828	4.828915	0.788462
0	Logistic Regreesion	76.923077	88.655914	4.107259	0.769231
1	SVM	75.824176	91.247312	4.552897	0.738462
4	Random Forest	72.527473	92.559140	5.776266	0.680769

Gambar 7. Hasil Evaluasi Menggunakan 9 Fitur Relevan

Berdasarkan Gambar 7, menunjukkan hasil evaluasi menggunakan 9 fitur relevan dari Chi2 dengan akurasi tertinggi didapatkan oleh algoritma KNN dan NB sebesar 78.02%. Jika dibandingkan dengan semua fitur, hasil evaluasi dengan 9 fitur ini menunjukkan peningkatan akurasi sebesar 2.2%.

	Model	Accuracy	Precision	Recall	F1
2	KNeighbors	79.120879	92.182796	4.442838	0.807692
3	GaussianNB	79.120879	87.655914	5.758825	0.807692
0	Logistic Regreesion	78.021978	88.967742	5.037999	0.788462
1	SVM	75.824176	90.913978	4.734908	0.738462
4	Random Forest	73.626374	93.204301	4.417719	0.688462

Gambar 8. Hasil Evaluasi Menggunakan 8 Fitur Relevan

Gambar 8 menunjukkan performa dari kelima model menggunakan 8 fitur relevan dari Chi2. Algoritma yang memperoleh akurasi tertinggi yaitu KNN dan NB sebesar 79.12%. Penggunaan 8 fitur relevan ini menunjukkan peningkatan sebesar 3.3% dari pada menggunakan semua fitur.

	Model	Accuracy	Precision	Recall	F1
3	GaussianNB	82.417582	88.623656	6.028951	0.842308
2	KNeighbors	81.318681	89.591398	4.334494	0.823077
1	SVM	80.219780	90.279570	4.771141	0.826923
0	Logistic Regreesion	80.219780	88.311828	5.057711	0.815385
4	Random Forest	76.923077	90.935484	5.714131	0.734615

Gambar 9. Hasil Evaluasi Menggunakan 7 Fitur Relevan

Berdasarkan Gambar 9 diketahui algoritma NB memperoleh nilai akurasi tertinggi yaitu 82.41% dengan menggunakan 7 fitur relevan dari *Chi-square*. Peningkatan yang diperoleh sebanyak 6.59% bila dibandingkan dengan semua fitur.

Penggunaan 7 fitur ini memperoleh peningkatan yang paling banyak bila dibandingkan dengan

DOI: https://doi.org/10.33020/saintekom.v15i1.815

pengujian menggunakan 10, 9 dan 8 fitur relevan. Adapun 7 fitur relevan ini terdiri dari *oldpeak, thalach, ca, exang, cp, slope* dan *thal*. Pengurangan fitur ini dapat mempercepat proses kinerja model. Dengan menggunakan 7 fitur relevan ini menunjukkan bahwa sistem yang dihasilkan lebih akurat dan efisien.

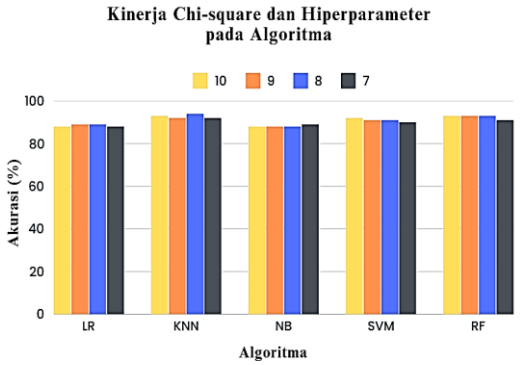
Tabel 2. Hasil Evaluasi Menggunakan *Chi-square* dan *Hyperparameter GridSearchCV*

Model	Jumlah Fitur			
	10(%)	9(%)	8(%)	7(%)
KNN	92.87	92.22	**93.51**	91.55
NB	88.00	88.31	87.66	88.62
LR	87.69	88.98	88.97	88.31
SVM	91.58	91.25	90.91	90.28
RF	92.88	92.56	93.20	90.94

Berdasarkan Tabel 2 menunjukkan hasil evaluasi model berdasarkan *Chi-square* dan *hyperparameter GridSearchCV* dengan *cross validation* = 10 dan *random state* = 42, dengan menggunakan 8 fitur relevan diperoleh algoritma KNN lebih unggul dibanding algoritma lainnya.

Algoritma KNN bekerja dengan baik dan mendapatkan akurasi tertinggi sebesar 93.51% dan disusul dengan algoritma RF dengan akurasi 93.20%.

3.2. Diskusi



Gambar 10. Kinerja *Chi-square* dan *Hyperparameter* pada Algoritma

Pada Gambar 10 merupakan hasil dari dioptimasi menggunakan *hyperparameter GridSearchCV* dengan *cross validation* = 10 dan *random state* = 42. Berdasarkan Gambar 10 menunjukkan akurasi tertinggi yang diperoleh pada penelitian ini berdasarkan 8 fitur relevan dari *feature selection* Chi2 yaitu sebesar 93.51%. Adapun 8 fitur relevan tersebut meliputi *oldpeak, thalach, ca, exang, cp, slope, thal* dan *sex*.

Tabel 3. Perbandingan Penelitian Sebelumnya

Dataset	Peneliti	Algoritma	Feature Selection	Accuracy
(Reddy et al., 2021)	Dataset Cleveland	SMO (Unggul), NB, LR, KNN, AdaBoostM1+DS, AdaBoostM1+LR,	Chi-square (Unggul), CFS dan Reflief.	86.46%

DOI: https://doi.org/10.33020/saintekom.v15i1.815

		Bagging+REPTree, Bagging+LR, Jrip dan RF.		
(Khurana et al., 2021)	Dataset Cleveland	SVM (Unggul), NB, DT, LR, RF dan KNN.	Chi-Square (Unggul), Gain Ratio, Information Gain, One-R and Relief.	83.41%
(Sarra et al., 2022)	Dataset Cleveland dan Dataset Statlog	SVM	Chi-square	89.40% (Cleveland)
Penelitian kami	Dataset Cleveland	KNN (Unggul), LR, NB, SVM, dan RF.	Chi-square	93.51%

Berdasarkan Tabel 3, beberapa penelitian sebelumnya dibandingkan dengan penelitian ini. Adapun pada penelitian (Reddy et al., 2021) memprediksi penyakit jantung dengan menggunakan dataset yang sama dengan penelitian ini dan diperoleh nilai akurasi tertinggi dari SMO sebesar 86.46% dengan 11 fitur dari Chi2. Bila dibandingkan dengan penelitian Reddy dkk., penelitian ini memperoleh peningkatan tidak hanya dari hasil akurasi melainkan juga dari jumlah fitur yang digunakan.

Pada penelitian (Khurana et al., 2021) menggunakan 9 fitur sehingga memperoleh hasil akurasi tertinggi dari SVM sebesar 83,41%. Hasil tersebut diperoleh dari *feature selection Chi-square* dan *Information Gain* yang lebih unggul dibandingkan dengan *feature selection* lainnya. Pada penelitian ini jika dibandingkan dengan penelitian Khurana dkk., maka penelitian ini lebih unggul baik dari segi akurasi maupun jumlah fitur relevan yang digunakan lebih sedikit.

Sedangkan pada penelitian (Sarra et al., 2022) juga melakukan prediksi penyakit jantung menggunakan dataset Cleveland UCI dan dataset Statlog, penelitian tersebut memperoleh akurasi dari dataset Cleveland UCI sebesar 89.40% dengan menggunakan 6 fitur dari *Chi-square*. Bila dibandingkan penelitian ini memberikan hasil yang lebih akurat daripada penelitian Sarra dkk.

4. KESIMPULAN

Penelitian ini menggunakan *feature selection Chi-square* dalam

DOI: https://doi.org/10.33020/saintekom.v15i1.815

mengklasifikasikan penyakit jantung dengan membandingkan kinerja dari algoritma klasifikasi LR, KNN, NB, SVM dan RF yang sudah dioptimasi menggunakan *GridSearchCV*. Dataset *heart disease* yang digunakan diambil dari UCI *Repository Machine Learning*, terdiri dari 303 baris dan 14 fitur. Dataset tersebut kemudian dibagi menjadi dua yaitu data *training* sebanyak 70% dan data *testing* sebanyak 30%. Dari hasil penelitian diketahui bahwa tidak selamanya semua fitur bisa menghasilkan hasil yang akurat, melainkan dengan mengurangi fitur-fitur yang tidak relevan bisa menghasilkan sistem yang akurat dan efisien. Berdasarkan hasil penelitian diperoleh algoritma KNN lebih unggul dibandingkan dengan algoritma lainnya yaitu dengan akurasi 93.51% dari 8 fitur relevan, dan disusul oleh algoritma *Random Forest* dengan akurasi 93.20% dari 8 fitur relevan juga. Adapun 8 fitur relevan tersebut adalah *oldpeak, thalach, ca, exang, cp, slope, thal* dan *sex*.

Untuk penelitian yang akan datang disarankan untuk menerapkan *feature selection Chi-square* ini pada dataset yang lebih besar lagi, bila perlu pada data real time sehingga bisa memberikan wawasan yang terkini. Tidak hanya itu, penelitian berikutnya diharapkan lebih mengeksplor setiap fitur dan *hyperparameter* yang digunakan.

DAFTAR PUSTAKA

Adiatma, B. C. L., Utami, E., & Hartanto, A. D. (2021). Pengenalan Ekspresi Wajah Menggunakan Deep Convolutional Neural Network. *EXPLORE*, *11*(2), 75. https://doi.org/10.35200/explore.v11i2.478

Albert, A. J., Murugan, R., & Sripriya, T. (2023). Diagnosis of heart disease using oversampling methods and decision tree classifier in cardiology. *Research on Biomedical Engineering*, *39*(1), 99–113. https://doi.org/10.1007/s42600-022-00253-9

Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective Heart-Disease Prediction by Using Hybrid Machine Learning Technique. *MDPI*, 1670–1675. https://doi.org/10.1109/ICCPCT58313.2023.10245785

Biswas, N., Ali, M. M., Rahaman, M. A., Islam, M., Mia, M. R., Azam, S., Ahmed, K., Bui, F. M., Al-Zahrani, F. A., & Moni, M. A. (2023). Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques. *Hindawi BioMed Research International*, *2023*. https://doi.org/10.1155/2023/6864343

Bujang, S. D. A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma,

DOI: https://doi.org/10.33020/saintekom.v15i1.815

E., Fujita, H., & Ghani, N. A. M. (2021). Multiclass Prediction Model for Student Grade Prediction Using Machine Learning. *IEEE Access*, *9*, 95608–95621. https://doi.org/10.1109/ACCESS.2021.3093563

Chandrasekhar, N., & Peddakrishna, S. (2023). Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization. *MDPI*, *11*(4). https://doi.org/10.3390/pr11041210

Claesen, M., & De Moor, B. (2015). Hyperparameter Search in Machine Learning. *ArXiv*, 10–14. http://arxiv.org/abs/1502.02127

Escalante, H. J. (2005). A comparison of outlier detection algorithms for machine learning. *Programming and Computer Software*.

Estetikha, A. K. A., Gutama, D. H., Pradana, M. G., & Wijaya, D. P. (2021). Comparison of K-Means Clustering & Logistic Regression on University data to differentiate between Public and Private University. *IJIIS: International Journal of Informatics and Information Systems*, *4*(1), 21–29. https://doi.org/10.47738/ijiis.v4i1.74

G, A., Ganesh, B., Ganesh, A., Srinivas, C., Dhanraj, & Mensinkal, K. (2022). Logistic regression technique for prediction of cardiovascular disease. *Global Transitions Proceedings*, *3*(1), 127–130. https://doi.org/10.1016/j.gltp.2022.04.008

Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. M. J. M., Ignatious, E., Shultana, S., Beeravolu, A. R., & De Boer, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques. *IEEE Access*, *9*, 19304–19326. https://doi.org/10.1109/ACCESS.2021.3053759

Jusia, P. A., Rahim, A., Yani, H., & Jasmir, J. (2024). Improving Performance of KNN and C4.5 using Particle Swarm Optimization in Classification of Heart Diseases. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *8*(3), 333–339. https://doi.org/10.29207/resti.v8i3.5710

Khairi, A., Ghozali, A. F., & Hidayah, A. D. N. (2021). Implementasi K-Nearest Neighbor (KNN) untuk Mengklasifikasi Masyarakat Pra-Sejahtera Desa Sapikerep Kecamatan Sukapura. *TRILOGI: Jurnal Ilmu Teknologi, Kesehatan, Dan Humaniora*, *2*(3), 319–323. https://doi.org/10.33650/trilogi.v2i3.2878

Khan, A., Qureshi, M., Daniyal, M., & Tawiah, K. (2023). A Novel Study on Machine Learning Algorithm-Based Cardiovascular Disease Prediction. *Health & Social Care in the Community*, *2023*(Cvd), 1–10. https://doi.org/10.1155/2023/1406060

Khurana, P., Sharma, S., & Goyal, A. (2021). Heart Disease Diagnosis: Performance Evaluation of Supervised Machine Learning and Feature Selection Techniques. *Proceedings of the 8th International Conference on Signal Processing and Integrated Networks, SPIN 2021*, *August*, 510–

DOI: https://doi.org/10.33020/saintekom.v15i1.815

515. https://doi.org/10.1109/SPIN52536.2021.9565963

Kibria, H. B., & Matin, A. (2022). The severity prediction of the binary and multi-class cardiovascular disease − A machine learning-based fusion approach. *ArXiv*, *98*, 107672. https://doi.org/10.1016/j.compbiolchem.2022.107672

Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare. *IEEE Access*, *8*(Ml), 107562–107582. https://doi.org/10.1109/ACCESS.2020.3001149

Narayanan, & Jayashree. (2024). Implementation of Efficient Machine Learning Techniques for Prediction of Cardiac Disease using SMOTE. *Elsevier*, *233*(2023), 558–569. https://doi.org/10.1016/j.procs.2024.03.245

Ogundepo, E. A., & Yahya, W. B. (2023). Performance analysis of supervised classification models on heart disease prediction. *Innovations in Systems and Software Engineering*, *19*(1), 129–144. https://doi.org/10.1007/s11334-022-00524-9

Ozcan, M., & Peker, S. (2023). A classification and regression tree algorithm for heart disease modeling and prediction. *Healthcare Analytics*, *3*(November 2022), 100130. https://doi.org/10.1016/j.health.2022.100130

Plackett, R. L. (1984). Karl Pearson and the Chi-squared Test. *International Statistical Institute*, *64*(1), 50–53. https://doi.org/10.47316/cajmhe.2024.5.1.05

Poslavskaya, E., & Korolev, A. (2023). *Encoding categorical data: Is there yet anything "hotter" than one-hot encoding?*

Rao, P. V., & Srivastava, K. K. (2024). Extraction and Feature Selection for Precise Cardiovascular Disease Classification. *International Journal for Multidimensional Research Perspectives*, *2*(7), 79–87. https://doi.org/10.61877/ijmrp.v2i7.172

Reddy, K. V. V., Elamvazuthi, I., Aziz, A. A., Paramasivam, S., Chua, H. N., & Pranavanand, S. (2021). Heart disease risk prediction using machine learning classifiers with attribute evaluators. *MDPI*, *11*(18). https://doi.org/10.3390/app11188352

Sarra, R. R., Dinar, A. M., Mohammed, M. A., & Abdulkareem, K. H. (2022). Enhanced Heart Disease Prediction Based on Machine Learning and χ2 Statistical Optimal Feature Selection Model. *MDPI*, *6*(5). https://doi.org/10.3390/designs6050087

Shaon, M. S. H., Karim, T., Shakil, M. S., & Hasan, M. Z. (2024). A comparative study of machine learning models with LASSO and SHAP feature selection for breast cancer prediction. *Elsevier*, *6*(February 2023), 100353. https://doi.org/10.1016/j.health.2024.100353

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance

DOI: https://doi.org/10.33020/saintekom.v15i1.815

measures for classification tasks. *Elsevier*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Spencer, R., Thabtah, F., Abdelhamid, N., & Thompson, M. (2020). Exploring feature selection and classification methods for predicting heart disease. *Digital Health*, *6*, 1–10. https://doi.org/10.1177/2055207620914777

Sushma, S. J., Assegie, T. A., Vinutha, D. C., & Padmashree, S. (2021). An improved feature selection approach for chronic heart disease detection. *Bulletin of Electrical Engineering and Informatics*, *10*(6), 3501–3506. https://doi.org/10.11591/eei.v10i6.3001

World Health Organization. (2021). *Cardiovascular diseases (CVDs)*. World Health Organixation. https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Yahaya, L., Oye, N. D., & Adamu, A. (2020). Performance Analysis of Some Se-Lected Machine Learning Algo-Rithms on Heart Disease Predic-Tion Using the Noble Uci Datasets. *International Journal of Engineering Applied Sciences and Technology*, *5*(1), 36–46. https://doi.org/10.33564/ijeast.2020.v05i01.006

Yulianto, S. P. R., Fanani, A. Z., Affandy, A., & Aziz, M. I. (2024). Analisis Metode Smoote pada Klasifikasi Penyakit Jantung Berbasis Random Forest Tree. *Jurnal Media Informatika Budidarma*, *8*(3), 1460. https://doi.org/10.30865/mib.v8i3.7712

DOI: https://doi.org/10.33020/saintekom.v15i1.815